N91-22793

PRELIMINARY RESULTS OF INVESTIGATIONS INTO THE USE OF ARTIFICIAL NEURAL NETWORKS FOR DISCRIMINATING GAS CHROMATOGRAPH MASS SPECTRA OF REMOTE SAMPLES

Harold A. Geller
Institute for Computational Sciences and Informatics
Department of Physics, George Mason University
Fairfax, Virginia
and Research and Data Systems Corporation
Greenbelt, Maryland
Eugene Norris
Department of Computer Science, George Mason University
Fairfax, Virginia
Archibald Warnock III
ST Systems Corporation
Lanham, Maryland

ABSTRACT

Neural networks trained using mass spectra data from the National Institute of Standards and Technology (NIST) are studied. The investigations also included sample data from the gas chromatograph mass spectrometer (GCMS) instrument aboard the Viking Lander, obtained from the National Space Science Data Center. We describe here the work performed to date and the preliminary results from the training and testing of neural networks. These preliminary results are presented for the purpose of determining the viability of applying artificial neural networks in discriminating mass spectra samples from remote instrumentation such as the Mars Rover Sample Return Mission and the Cassini Probe.

INTRODUCTION

Artificial neural networks are a form of artificial intelligence that may be useful in categorizing data, particularly data that have recognizable patterns as the basis for discriminating sets within a larger group. Successful applications include optical character and speech recognition. A properly trained neural network should be able to discriminate assays using mass spectrometry in conjunction with gas chromatography. Such sample analyses are being planned for automated instrument missions to Mars and Titan. There exists a requirement to develop a light-weight, rapid capability to discriminate sample analyses and provide a first order of magnitude recommendation for further Earth-based analysis for those craft which will return sample analyses via downlinks or actual return vehicles.

The Mars Rover Sample Return vehicle will likely require its own ability to choose the samples that are returned to Earth. The Cassini probe instruments may suffer from limited transmission bandwidth thus requiring remote decisions as to what sample analyses should be transmitted back to Earth for further investigation. This preliminary investigation was undertaken to discover the feasibility of using artificial neural networks in the analyses of data and the decision making for determining the best prospects for further analysis.

Chromatography itself is the separation technique by which a sample is distributed between two phases and is resolved based upon its differential adsorption between the two phases or media (Khandpur,1981). In gas chromatography, a gas is used to transport the sample through the chromatographic column. The detector at the end of the column is used to determine

the individual peaks that develop based on the time it takes for the constituents to pass through the column. Differences depend on the molecular adhesion to the column's own molecular components.

Gas chromatography is an excellent separation technique but suffers from poor identification of the constituents. Detection techniques are often used after separation in a chromatographic column to positively identify constituents of interest. The technique used to distinguish elements by their different mass to charge ratio of the ionic state is called mass spectrometry (Khandpur,1981). The sample is ionized and passed through a chamber with specific electric and magnetic fields. Detectors are placed so that peaks occur where ions are found and these peaks can tell the investigator the element which constituted the original sample (Khandpur,1981 and Message, 1984). When a mass spectrometer is used in conjunction with a gas chromatograph it is referred to as a single instrument, a gas chromatograph mass spectrometer (GCMS).

METHODOLOGY

The Viking data is available from the National Space Science Data Center (NSSDC), located at Goddard Space Flight Center (GSFC) in Greenbelt, Maryland. The data at NSSDC, however, is currently only available from magnetic tapes which are binary-formatted for an IBM 1800.

In attempts to locate the data in a more easily read format by a neural network, such as ASCII, contact was made with numerous members of the original MIT GCMS investigation team. The GCMS data from the Viking Lander was never converted to any alternative format which led to attempts to work with the original data as provided by NSSDC.

The original six tapes sent to NSSDC were received from NSSDC by this investigation team as a single 1600 bpi tape. It contained 6 files corresponding to the original 6 tapes of data. Only the third and sixth files contains processed data from Viking Land 1 and Viking Lander 2 respectively.

There is an ongoing effort to convert this data to ASCII for the neural network, however at the same time, a limited amount of Viking Lander data was read from the microfilm archive and used for testing purposes.

For training the neural network we acquired Version 3.0 of the National Institute of Standards and Technology (NIST) PC database of electron ionization mass spectra (NIST, 1990). The search software provided with the PC mass spectra database could be used as a basis for comparison with the performance of trained neural networks. We describe two search functions which require the user to specify the peak abundances of an unknown compound. One search is by abundances of major peaks ("M" search) and the other is a search by the presence of any specified peak ("A" search).

The "M" search locates spectra in the database that display peak abundance characteristics for the largest peaks. These peaks searched are exactly the same as those of the spectrum of the unknown. This search only retrieves spectra for which the relative ordering of peak abundances that exactly match those specified.

The "A" search, adapted from a spectral search developed by Heller (NIST, 1990) requires the user to enter the mass of a peak that appears in the unknown spectra and a relative abundance range. This search allows for entering up to 10 peaks with their respective masses and abundances.

Both NIST searches terminate with a list of structures and names of compounds which are plausible matches for the unknown spectra. It is then left to the investigator to examine those spectra to determine the most likely constituents.

TRAINING ISSUES

The back-propagation (BP) algorithm was selected as most the promising architecture for detecting the various organic molecules and their fragments. This choice was based on the need for a supervised pattern recognition algorithm that can be easily modified to accommodate different data collection scenarios. It is possible that the final architecture to be adopted for a GCMS application will benefit from having several BP modules trained independently to detect mass spectral fragments.

All pattern recognition systems require some form of training. In traditional pattern recognition systems, training consists of statistical characterization of the data using mathematical representations. For the BP algorithm, however, training consists of repeated presentations of data samples selected to be typical of the data that will be encountered in the operational setting. Each sample (or fact) is a vector of measurements from the data which is accompanied by a desired characterization. The network responds to the error signals and attempts to find a non-parametric characterization of the data set that is consistent with the training data set.

The selection of a training data set is a very important consideration. The number of samples in the training set should ideally be based upon the total number of spectra available, the *a priori* probabilities of occurrence for each spectra within Viking data, and the ability of the original data extraction techniques to present a pure spectrum.

The first criterion was bounded by the total number of spectra in the NIST data base. The second criterion could not be determined for this first analysis. The third criterion was difficult to quantify, and based solely on investigator experience. Sampling theory may ultimately provide a modicum of guidance.

The mass spectrometry technique causes the development of molecular fragments. This knowledge was used by the developers of DENDRAL, who decided to have the expert system search not for whole spectra but for fragments of spectra. It then developed a list of molecules which could form the given fragments and use additional information for narrowing the possibilities to a single or chosen few. This fragmentary analysis approach is the one often used by mass spectra analysis experts, but was too complex for this preliminary neural network implementation. The approach taken was to train the network using complete mass spectra and discover if the network could then guess the molecular make-up of the unknown.

In order to assess the performance of an artificial neural network it is necessary to apply it to one or more test sets of data having known categories (ground truth in neural network literature). If the system performance on the training and test sets is similar, the network should exhibit good generalization properties and its behavior, for example on a spacecraft on a distant planet, should produce few surprises.

On the other hand, if the performance of the network on the test set is much worse that that of the training set, then the classifier design should be re-examined. In our study, the only case available for testing was the Viking data set. The Viking data sets themselves might have been partitioned into training and testing sets. A shortage of training data may reduce the robustness of the network whereas a shortage of testing data may reduce the confidence in the measurement of the network's performance.

Viking Mission GCMS Datasets

On July 20th, 1976, the first Viking Lander descended upon the soil of Mars (Soffen, 1977 and Snyder, 1977). The x-ray fluorescence spectrometer, which was on board to discover the inorganic composition (Clark et al, 1977) determined the Martian soil to be composed of between 15-30 percent silicon, 12-16 percent iron, 3-8 percent calcium and 2-7 percent aluminum (Clark et al, 1977 and Toulmin et al, 1977).

The gas chromatograph mass spectrometer was said to have given an indication of a little water but no organic compounds (Biemann et al, 1977). This conflicted with results from the biology experiments which indicated the existence of microbial life (Horowitz and Hobby, 1977, Levin and Straat, 1977, Oyama and Bordahl, 1977 and Klein, 1977).

All of the data transmitted by the two Viking Landers was ultimately deposited with the NSSDC. An example of a filtered mass spectra is presented in Figure 1. A sample of a mass spectrum from the Viking Lander is shown in Figure 2 illustrating substantial amount of filtering, and is just one stage where errors can be introduced in the analyses.

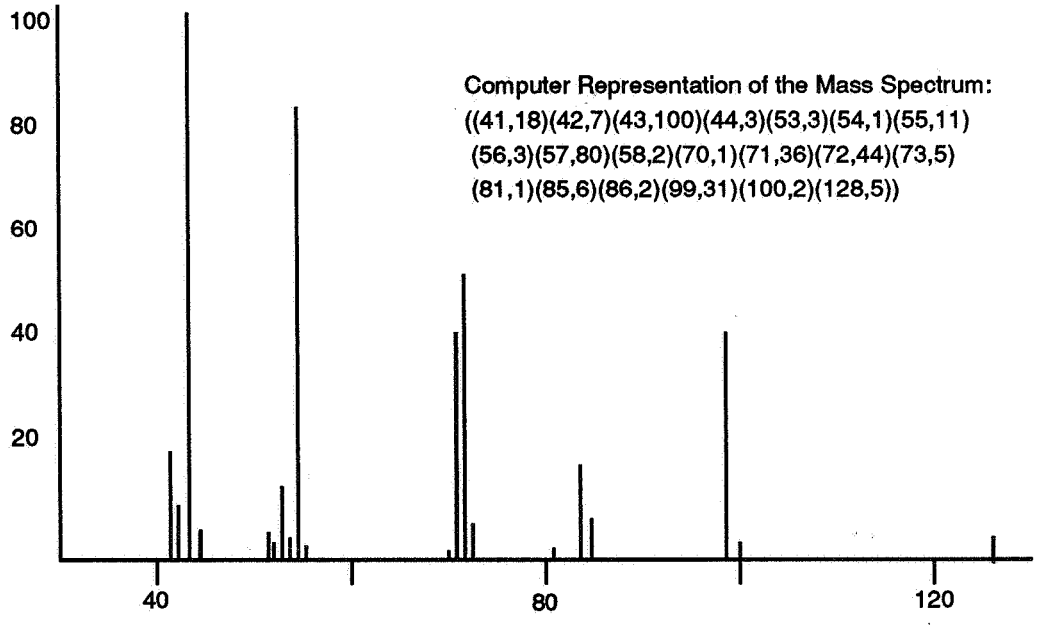

Figure 1 Mass Spectrum for 3-Octanone

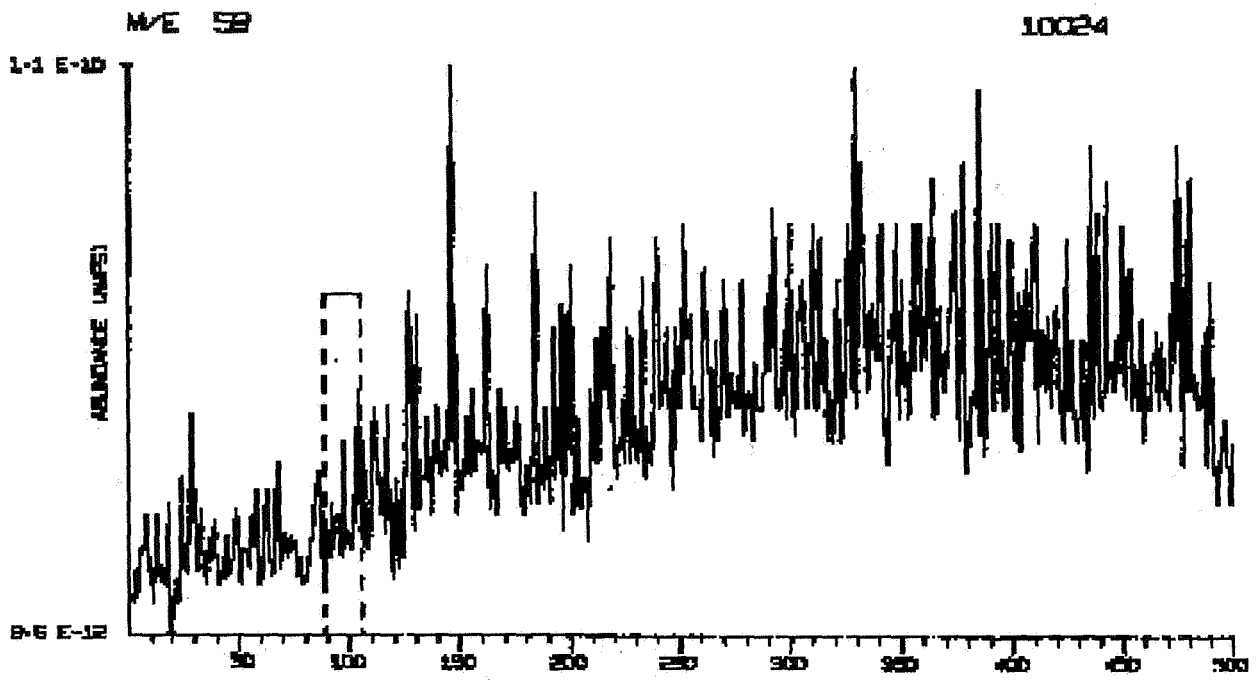

Figure 2 Viking Mass Spectra Results

RESULTS

The neural network simulation software used in this investigation was the commercial package Brainmaker (CSS, 1990). The first attempts at training the neural network were done using a strict pattern recognition approach to the problem. The spectra are classically represented using a format whereby a list of ordered pairs represents the mass-to-charge ratio and a relative abundance figure. For example, if the highest abundance of fragmentary material from the mass spectrometer has a mass-to-charge ratio of 35, its peak would be represented by the order pair, (35,99). NIST used a scale ranging from 0 to 999 (maximum peak of 999) to represent relative peak abundances, and this was the scale adopted for this investigation.

The first attempts at training a neural network were accomplished by using a pictorial representation of the mass spectra in which individual peaks of spectra were represented by "filled" pixels at the coordinates given by the ordered pair associated with the peak. These first attempts, although yielding a trained network, were disappointing because combinations of spectra would totally baffle the network.

A different approach in the visualization on screen and the format of the data was then attempted. The on-screen visualization developed for this subsequent series used a thermometer type representation of inputs, outputs and neuron firings. Output neurons corresponded to the molecules in the training set and the higher the output neuron value, the longer the bar graph representation.

Having decided that it was most prudent to limit the number of input neurons as much as possible, even pairs of numbers were too cumbersome for data entry. The final format used included a series of numbers corresponding to the abundance at the mass-to-charge ratio represented by the node. The nodes of the first layer were ordered to make a one-to-one correspondence to the mass-to-charge ratio value.

A subset of mass spectra from the NIST database was selected as the core training set by applying two criteria:

1) The chemicals in the training set should have a history of being discovered in mass spectra from sources other than Earth
2) The molecular weight of the compound should be less than 100 (to minimize the effort required for formatting the data)

One source of GCMS chemical compounds examined in extraterrestrial samples was derived from the analysis of organogenic compounds in Apollo 11, 12, and 14 lunar samples (Flory, et al, 1972). Additional compounds were derived from an analysis of amino acid precursors in lunar samples from Apollo 14 and earlier (Fox, et al, 1972).

A list of the molecules used to train the preliminary neural networks can be seen in Figure 3. First a neural network using mass spectra of the chosen chemicals from the NIST database was trained. The network architecture used for the training was a 3-layer network with 100 input neurons, 9 hidden-layer neurons, and 14 output neurons. Associated with the 9 hidden-layer neurons is 909 weights while there are 140 weights associated with the 14 output neurons.

The input neurons corresponds to a mass-to-charge ratio (1 to 100) and each output neuron corresponds to a single chemical compound (1 to 14). The learning parameters used for the training of the network were a learning rate of 1.0, a training tolerance of 0.1, a testing tolerance of 0.4, and a smoothing factor of 0.9.

This network required 8 minutes 3 seconds to be successfully trained on a PC compatible 386SX with no math co-processor. During this training period the neural network had examined all 14 mass spectra 288 times, yielding a total reading of some 4032 spectra (14 x 288).

To determine the usefulness of adding training data of other than individual compounds, a training set was developed which also included a combination of two spectra for identification. The combination included in the training set was 50% water and 50% serine. The reason for choosing this combination was that these compounds do not have any overlapping peaks

The training of the neural network proceeded using the same network parameters and architecture. This network took 5 minutes 22 seconds and read all spectra 179 times. Each spectral set contained 15 spectra (14 individual spectra and one combination) for a total of 2685 spectra read by the network.

In comparing the weight matrices associated with these networks, we discovered that the weights for many nodes converged to the same value, although the second net converged much quicker.

Investigations into the possible corruption of data by random noise across all m/e values were performed. Figure 3 represents the interpretation of a spectrum whose m/e values (1-100) were coupled with three digit pseudo-random generated numbers (0-999).

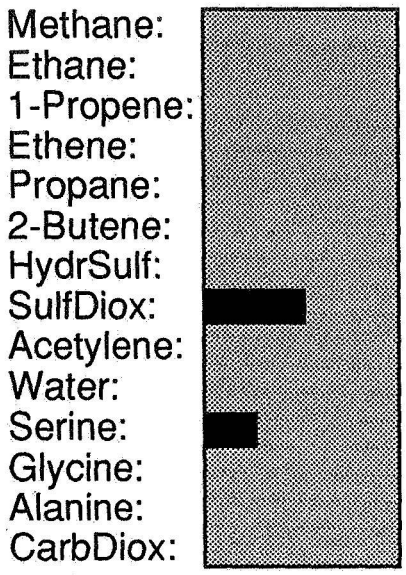

Figure 3. 3-Digit Random Values Throughout

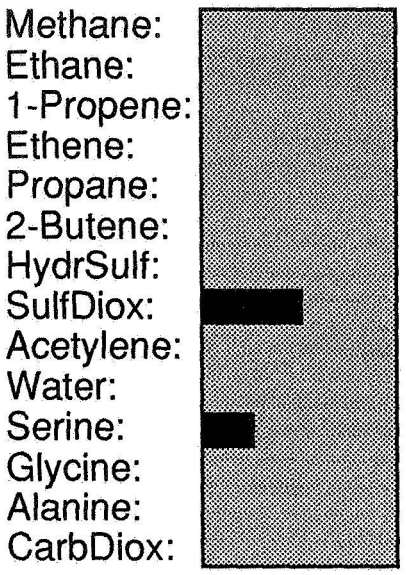

Figure 4. 3-Digit Random Values with Ethane Present

Figure 4 represents the interpretation of a spectrum which was developed by filling all m/e values with three digit pseudo-random generated numbers except at the positions of the peaks corresponding to the compound ethane.

Figure 5 represents the interpretation of a spectrum which was developed by filling all m/e values with two digit pseudo-random generated numbers except at the position of the peaks corresponding to the compound 1-propene.

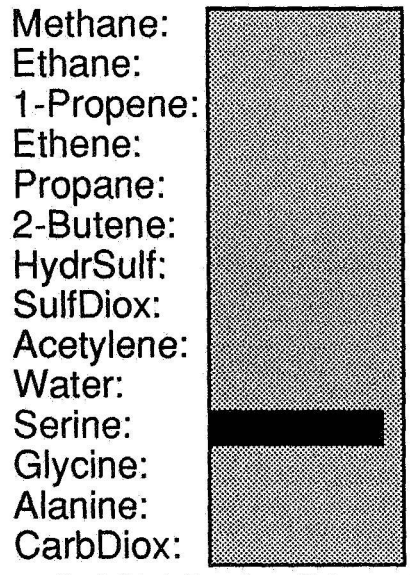

Figure 5. 2-Digit Random Values with 1-Propene Present

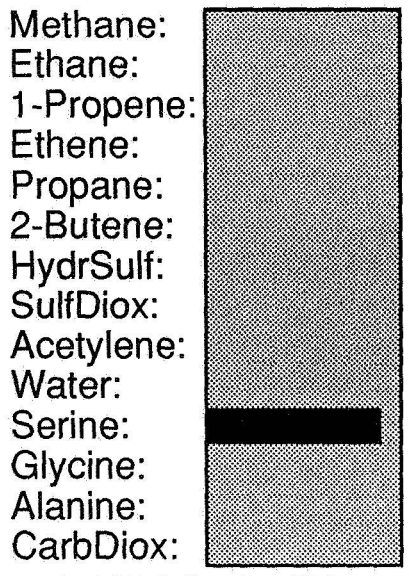

Figure 6. 1-Digit Random Values with Ethene Present

Figure 6 represents the interpretation of a spectrum which was developed by filling all m/e values with single digit pseudo-random generated numbers except at the position of the peaks corresponding to the compound ethene.

These results indicate the difficulty that the trained neural network has with the presence of random noise in the m/e peak values. This difficulty should be investigated further to determine a network's capability to identify spectra with noise similar to that which may be expected in actual remote sample analyses.

Three mass spectra from the Viking Lander were hand entered from microfilm copies of the spectra sent to NSSDC. One spectrum contained the raw data Spectra such as this represented, had to have peaks re-normalized in order to mask the peaks caused by carbon dioxide and water. When we presented the raw data to the trained network, it identified carbon dioxide (as the predominant constituent) and water (as a trace constituent) as depicted in Figure 7.

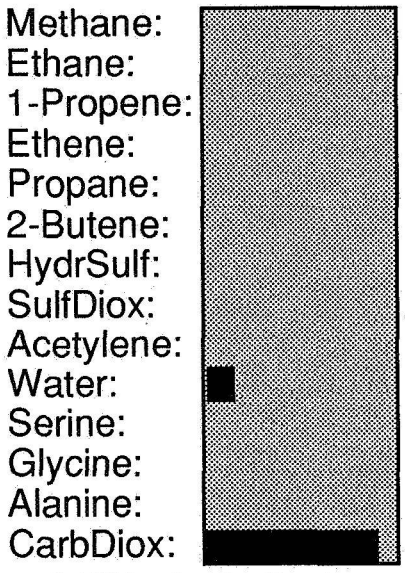

Figure 7. Viking Lander Spectrum with All m/e Peaks Present

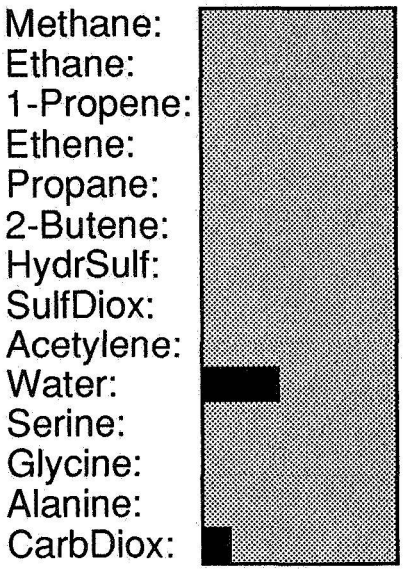

Figure 8. Tighter Tolerance Network of Same Spectrum (Figure 7)

Testing continued with a trained network set to a tighter training tolerance (0.050 vice 0.100). The second net did not do as well as the first in the identification of the constituents as can be seen in Figure 8.

Recall that the Viking data was entered by reading off values from a graph derived from a microfilm hardcopy whose error was determined to be plus or minus 1 m/e value . This error led the team to investigate the ability of the neural network when test spectra were corrupted in a similar manner.

First, the spectra for all compounds were shifted one m/e value down. That is, the first m/e ratio was dropped and substituted with the second, the second with the third and so forth until the 100th m/e value was filled with a 0 level peak value. Next, spectral peak values were shifted one m/e value upwards, in a similar manner as described.

Results from these shifted spectra were mixed. When tested with the spectra of the compounds down-shifted by one m/e value, the neural network was still able to positively

identify (within testing tolerance) 8 of the 14 compounds. However, when all spectra were up-shifted by one m/e value, the network identified correctly 6 of the 14 compounds.

CONCLUSIONS AND RECOMMENDATIONS

We have sought to demonstrate the feasibility of using artificial neural networks in the discrimination of GCMS samples for the purpose of a fast and simple means for choosing interesting samples to be further analyzed in a laboratory when such analysis is not available in-situ. To this end the investigators have demonstrated the following:

1) A neural network can be trained to identify individual mass spectra of chemical compounds

2) A neural network can identify molecules whose data has been corrupted by shifts in spectral peaks

3) A neural network trained with combinations of mass spectra can accomplish its training in a shorter timeframe than one with only individual mass spectra

4) A neural network trained for identifying individual mass spectra has difficulty in interpreting mass spectra with a large amount of randomly generated data throughout the spectra

This investigation has concentrated on the second portion of the GCMS, namely the mass spectrometry. However, information on the effluent could be used in the training of the neural network which would benefit the discrimination of the sample to a finer degree than without the GC data. Separation techniques themselves are prone to certain errors as well (Silverstein and Geller, 1974). The accuracy of the Viking Lander GCMS has at least one critic, and doubt of the sensitivity of the instrument lingers to this day (Levin and Straat, 1988).

We believe that using an artificial neural network in the analysis of complex chemical data sets may yet prove to be beneficial in the future unmanned exploration of Mars, Titan and other solar system bodies. Further investigations are warranted.

Acknowledgements: The authors would like to acknowledge all of those that played a part in the development of this investigation and the attempts to demonstrate a viable alternative for remote analyses. Those we wish to thank include Glenn Glover (SAIC), Michael Martin (NASA JPL), Klaus Biemann (MIT), Tom Ryan (SAIC), Mary Lawler-Covell (SAIC), Hasso Niemann (NASA GSFC), Ralph Post (STX) and the National Space Science Data Center. The authors also wish to acknowledge the financial sacrifices made by Harold Geller which contributed to the success of this effort.

REFERENCES

Biemann, K., Oro, J., Toulmin,P., Orgel, L., Nier, A., Anderson, D., Simmonds, P., Flory, D., Diaz, A., Rushneck, D., Biller, J. & Lafleur, A. (September 1977). The Search for Organic Substances and Inorganic Volatile Compounds in the Surface of Mars. *Journal of Geophysical Research* 82(28), 4641-4658.

Clark,B., Baird, A., Rose, H., Toulmin, P., Christian, R., Kelliher, W., Castro, A., Rowe, C., Keil, K. & Huss, G. (September 1977). The Viking X Ray Fluorescence Experiment: Analytical Methods and Early Results. *Journal of Geophysical Research* 82(28), 4577-4624.

CSS (December 1990) *Brainmaker User's Guide and Reference Manual 5th Edition* Grass Valley: California Scientific Software

Flory, D., Wikstrom, S., Gupta, S., Gibert, M., & Oro, J. (1972) Analysis of Organogenic Compounds in Apollo 11, 12, and 14 Samples. In Heymann, D (Ed.) *Proceedings of the Third Lunar Science Conference* (pp.2091-2108). Cambridge: MIT Press.

Fox, S., Harada, K., & Hare, P. (1972) Amino Acid Precursors in Lunar Fines from Apollo 14 and Earlier Missions. In Heymann, D (Ed.) *Proceedings of the Third Lunar Science Conference* (pp.2109-2129). Cambridge: MIT Press.

Horowitz, N., Hobby, G. and Hubbard, J. (September 1977) Viking on Mars: The Carbon Assimilation Experiments. *Journal of Geophysical Research* 82(28), 4659-4662.

Khandpur, R.S. (1981) *Handbook of Analytical Instruments*. Philadelphia, Tab Books, Inc.

Klein, H. (September 1977) The Viking Biological Investigation: General Aspects. *Journal of Geophysical Research* 82(28), 4677-4680.

Levin, G. & Straat, P. (September 1977) Recent Results From the Viking Labeled Release Experiment on Mars. *Journal of Geophysical Research* 82(28), 4663-4667.

Levin, G.V. & Straat, P. (1988) A Reappraisal of Life on Mars. In Reiber, D. (Ed.) *The NASA Mars Conference* Volume 71 (pp.187-207). San Diego: American Astronautical Society.

Message, G.M. (1984) *Practical Aspects of Gas Chromatography/Mass Spectrometry*. New York, John Wiley & Sons.

NIST (June 1990) *NIST/EPA/MSDC Mass Spectral Database PC Version 3.0 User's Guide* Gaithersburg: US Department of Commerce.

Oyama,V.I. & Bordahl,B.J. (September 1977) The Viking Gas Exchange Experiment Results From Chryse and Utopia Surface Samples. *Journal of Geophysical Research* 82(28), 4669-4676.

Silverstein,E. & Geller,H. (December 1974) Studies on the Nature of Non-Specific Staining in Nitro-Blue Tetrazolium Detection of Dehydrogenases in Polyacrylamide Gel Electrophoresis *Journal of Chromatography* 101(4), 327-337.

Snyder, C.W. (September 1977) The Missions of the Viking Orbiters. *Journal of Geophysical Research* 82(28), 3971-3983.

Soffen,G.A. (September 1977) The Viking Project. *Journal of Geophysical Research* 82(28), 3959-3970.

Toulmin, P., Baird, A., Clark, C., Keil, K., Rose, H., Christian, R.P., Evans, P.H. & Kelliher, W. Geochemical and Mineralogical Interpretation of Viking Inorganic Chemical Results. *Journal of Geophysical Research* 82(28), 4625-4634.